SEME at SemEval-2024 Task 2: Comparing Masked and Generative Language Models on Natural Language Inference for Clinical Trials

Mathilde Aguiar, Pierre Zweigenbaum, Nona Naderi

SEME at SemEval-2024 Task 2: Comparing Masked and Generative Language Models on Natural Language Inference for Clinical Trials

Mathilde Aguiar, Pierre Zweigenbaum, Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
{mathilde.aguiar, pierre.zweigenbaum, nona.naderi}@lisn.fr

Abstract

This paper describes our submission to Task 2 of SemEval-2024: Safe Biomedical Natural Language Inference for Clinical Trials. The Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) consists of a Textual Entailment (TE) task focused on the evaluation of the consistency and faithfulness of Natural Language Inference (NLI) models applied to Clinical Trial Reports (CTR). We test 2 distinct approaches, one based on finetuning and ensembling Masked Language Models and the other based on prompting Large Language Models using templates, in particular, using Chain-Of-Thought and Contrastive Chain-Of-Thought. Prompting Flan-T5-large in a 2-shot setting leads to our best system that achieves 0.57 F1 score, 0.64 Faithfulness, and 0.56 Consistency.

1 Introduction

The digitization of medical documents allows the development of tools using various NLP techniques. In the case of Clinical Trial Reports (CTR), these tools can facilitate recruiting patients to participate in a trial or help researchers keep up to date with the literature. Natural Language Inference (NLI) is particularly useful in detecting the relationship between a CTR and a statement. For instance, it can be used for patient-trial matching.

Task 2 of SemEval 2024 defines a Textual Entailment (TE) task applied to English breast cancer CTRs. A submitted system must perform a binary classification based on a CTR and a given statement, using the labels *entailment* or *contradiction*. In addition to the traditional F1-measure for Textual Entailment, the submitted systems are evaluated on 2 strong metrics: Faithfulness and Consistency.

In this paper, we first introduce the task and some related work in Sec. 2. Sec. 3 describes our proposed approaches, while Sec. 4 gives further details about the experimental setup. Sec. 5 presents the results and comparative analysis of methods, and Sec. 6 sums up our work done and provides ideas for future work.

2 Background

2.1 Corpus and task description

The NLI4CT (Jullien et al., 2024) corpus consists of a collection of breast cancer Clinical Trial Reports (CTR) taken from clinicaltrials.gov. The documents are exclusively written in English. These CTRs are structured with the following sections: *Intervention* section describes what treatment is going to be applied during the trial. *Eligibility* section consists of a set of inclusion and exclusion criteria that a test subject must comply with. *Results* section displays the outcome measures. Finally, *Adverse Events* section describes the side effects and symptoms observed during the trial. In NLI4CT there are two types of instances: *single*, where only 1 CTR is involved to perform the inference, and *comparison* where 2 CTRs need to be compared.

The task's objective is to perform Natural Language Inference on these clinical trials. A premise consists of a section of a CTR (or two CTRs if it is a comparison), and a statement is a single sentence. The model should predict whether the premise entails or contradicts the statement. To tackle the NLI4CT task, the model must perform several kinds of inference, such as quantitative, common-sense, and medical reasoning (see Fig. 2). The inference relationship can be predicted using the evidence, sentences where clues are contained, that are in one of the sections of a CTR. Evidence is provided only in the development and training sets. The dataset is balanced with half of the instances labeled as *entailment* and the other half as *contradiction* in the train and development subsets.

2.2 Related work

A previous edition of the NLI4CT task was run as SemEval 2023 Task 7 (Jullien et al., 2023a). It was composed of 2 subtasks: an NLI classification task and an information retrieval task of evidence selection to support the predicted label. The training and development sets were the same as the present edition. For the first subtask, the task overview paper (Jullien et al., 2023b) reports both generative and discriminative approaches for the submitted systems. Over the past few years, we have seen the fast-paced development of Large Language Models (LLMs) and their increased capabilities in addressing both generative and discriminative tasks. Even general-domain LLMs like Flan-T5-xxl in Kanakarajan and Sankarasubbu (2023) and GPT-3.5 in Pahwa and Pahwa (2023) have been achieving competitive performance on domain-specific tasks for the 2023 edition of the NLI4CT task.

3 System overview

To address the NLI4CT task, we tested 2 main approaches: the first uses Pretrained Masked Language Models (MLM), and the second uses generative Large Language Models. We wanted to compare the ability of these two kinds of architectures to solve the same task, in particular in terms of consistency and faithfulness.

3.1 Finetuning pretrained masked language models

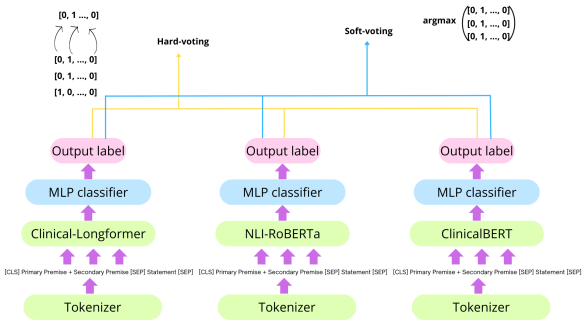

Figure 1: MLM ensemble architecture overview.

Our first system is based on finetuning and ensembling multiple MLMs on the task data (see an example in Fig. 1). We first finetune each model using the train and development splits of NLI4CT. We evaluate each finetuned model on the test set. We perform experiments with two ensembling methods: hard-voting and soft-voting. The hard-voting method consists of selecting the label y that gets the majority of votes across the predictions of each model j, defined as follows:

$$\tilde{y} = \underset{y}{\text{argmax}} \sum_{j=1}^{N} \mathbb{1}(\tilde{y}_j = y)$$

Soft-voting is computed by using the argmax of probabilities P_j from each model j for a given label y:

$$\tilde{y} = \underset{y}{\text{argmax}} \sum_{j=1}^{N} P_j(y)$$

3.2 Prompting generative large language models

We designed a set of prompts that rely on the following techniques:

1. A simple prompt instructing the model to perform Textual Entailment, giving the statement and a premise composed of the whole section where the evidence comes from. We took inspiration from the instruction templates found in the Flan-Muffin dataset[1] that (Lou et al., 2024) used to instruction-tune the Flan-T5 models (Chung et al., 2022). The template starts with optional demonstrations that instantiate this prompt with n training or development examples in n-shot settings:

[Demonstrations] [Premise] [Statement] Based on this premise, is the hypothesis true? OPTIONS: -'Yes' -'No'

2. Using the concept of Chain-Of-Thought (Wei et al., 2022) that decomposes the reasoning behind a given example; we insert the premise sentences that are the actual evidence used to infer an entailment or a contradiction in the demonstrations. See C.2 for a detailed example.

3. We tested the related Contrastive Chain-Of-Thought (CCOT) (Chia et al., 2023) technique that gives both one correct and one incorrect explanation in addition to the original template. In our case, we inserted premise sentences that were not actual evidence. See C.3 for an example. CCOT is inspired by how humans learn from positive and negative examples and aims to reduce reasoning errors by indicating what mistakes to avoid.

For the demonstrations, we tried three few-shot settings: zero-shot (ZS: no demonstration, only for the first template), 1-shot, and 2-shot. See Appendix C for detailed examples of the prompts.

[1] https://huggingface.co/datasets/causal-lm/flan-muffin

4 Experimental setup

4.1 Data pre-processing

We used the NLI4CT train and development splits published by BigBio on HuggingFace[2] and enriched them with new columns: primary and secondary evidence and premises from the JSON files provided by the organizers. We used this dataset to build our prompts (see Sec. 3.2). We shuffled the train and dev sets and selected random instances to include as demonstrations in our 1 and 2-shot settings.

4.1.1 Ensembling MLMs

We used Masked Language Models that are pretrained on general domain data or clinical data. For the general domain, we selected NLI-RoBERTa[3] (Reimers and Gurevych, 2019) from Sentence Transformers, which has been previously finetuned for NLI using SICK (Marelli et al., 2014) and STS benchmark (Cer et al., 2017). For the clinical pretrained models we use Clinical-Longformer[4] (Li et al., 2023), which can handle a context window up to 4096 tokens, and ClinicalBERT[5] (Wang et al., 2023) which has been pretrained on Electronic Health Records. We used Optuna (Akiba et al., 2019) for hyperparameter search and set our final configuration with a learning rate of $5e^{-5}$ using the AdamW (Loshchilov and Hutter, 2017) optimizer, a batch size of 64 and finetuned the models for 4 epochs. Ensembles of the same model used a different random seed when training each instance. We used 4 NVIDIA Tesla V100 with 32 GB of RAM with a training and inference time varying from 3 to 6.5 hours. A more detailed analysis of the training cost can be found in Appendix 6.

4.1.2 Prompting generative LLMs

We tested several Large Language Models (see Appendix F). We eventually chose Flan-T5-large[6] (Chung et al., 2022) for its ability to output answers that are easier to parse than the longer and more challenging answers that could be provided by Llama-2 (Touvron et al., 2023) or Mistral (Jiang et al., 2023). Flan-T5 has been pretrained on a mixture of 473 datasets covering 1,836 tasks. However, it has no biomedical or clinical pertaining. We rely on the HuggingFace framework for all experiments. We used the same computing setup as in the previous set of experiments. The codebase for all of our experiments is freely available.[7]

4.2 Evaluation

We evaluate our models using the following metrics. The F1 score of the *Entailment* class is measured on a control set of the gold test set which is the same as the NLI4CT 2023's test data. Faithfulness measures whether a model changes predictions when an 'entailing' statement is changed into a 'contradicting' statement. Consistency measures whether a model keeps its predictions when a statement is changed while preserving its relation to the premise. Both metrics are computed on a contrast set of the gold test set that has undergone perturbations (more details in Jullien et al. (2024)).

5 Results

5.1 Quantitative analysis

Under the username *math_agr*, our team ranked 27th for an F1 score of 0.57, 18th for Faithfulness of 0.64, and 25th for Consistency of 0.56. Tables 1–6 report the results of our experiments on the test set.

Single system	F1	Faithfulness	Consistency
Majority class	0.67	0.00	0.38
tf.idf (Jullien et al., 2024)	0.41	0.47	0.47
FZI-WIM	0.80	0.90	0.73
rezazzr	0.06	0.95	0.60
NYCU-NLP	0.78	0.92	0.81
a: NLI-RoBERTa	0.56	0.58	0.57
b: ClinicalBERT	0.00	**1.00**	**0.62**
c: Clinical-Longformer	**0.67**	0.00	0.38
Ensemble	s/h	s/h	s/h
(a+a+a)	0.57/0.57	0.58/0.54	0.57/0.56
(b+b+b)	0.56/0.63	0.37/0.16	0.47/0.43
(c+c+c)	**0.67**/0.64	0.00/0.09	0.38/0.40
d: (a+b+c)	0.55/0.57	0.45/0.40	0.52/0.52
(d) + Flan-T5-large	0.57 (h)	0.64 (h)	0.56 (h)

Table 1: F1 score, Faithfulness, and Consistency for single Masked Language Models then soft (s) and hard (h) ensembling. Ensembles such as (a+a+a) consist of 3 instances of the same model. Flan-T5-large is used in a 2S setting (see Tab. 2 below).

[2] https://huggingface.co/datasets/bigbio/sem_eval_2024_task_2

[3] https://huggingface.co/sentence-transformers/nli-roberta-base-v2

[4] https://huggingface.co/yikuan8/Clinical-Longformer

[5] https://huggingface.co/medicalai/ClinicalBERT

[6] https://huggingface.co/google/flan-t5-large

[7] https://github.com/MathildeAguiar/SemEval-2024-Task-2

Each model has different strengths and weaknesses across the three metrics in the MLM experiments. The single NLI-RoBERTa seems to be the most stable baseline despite its lack of pretraining on biomedical data. It has already been finetuned on general-domain NLI, and its sentence-level representation seems to boost its performance. The ensemble of 3 NLI-RoBERTa does not add enough diversity to improve its results. The single ClinicalBERT obtains an F1-score of 0.00: we observed that it always predicts the label *Contradiction*, which causes a precision and recall of 0.00. Faithfulness yields 1.00 because it is computed on instances of the contrast test set that are all labeled as *Contradiction*. The ensemble of 3 ClinicalBERT does not have this issue: some seeds led to better models. The single Clinical-Longformer obtains the best results in terms of F1-score but the worst on the other two metrics, especially on Faithfulness. It predicts almost exclusively *Entailment*, which leads to Faithfulness and Consistency complementary to ClinicalBERT's. The ensemble keeps the same issues. An ensemble (d) of the three single models could not improve the single NLI-RoBERTa. Adding Flan-T5's 2-shot predictions to the ensemble increased Faithfulness by 0.24 points but did not yield better F1. This did not improve either over Flan-T5 alone (see row 2S in Tab. 2).

Prompt	F1	Faithfulness	Consistency
ZS	0.56	0.57	0.55
1S	0.53	0.63	**0.57**
2S	**0.57**	0.64	0.56
1SCOT	0.39	0.70	0.53
2SCOT	0.43	0.69	0.51
1SCCOT	0.28	**0.85**	**0.57**
2SCCOT	0.24	0.81	0.56

Table 2: F1 score, Faithfulness, and Consistency for the LLM approach, using Flan-T5-large.

Prompting Flan-T5-large in few-shot mode performs as well as the fine-tuned NLI-RoBERTa. Increasing the number of demonstrations tends to improve the scores. This illustrates the usual trade-off between fine-tuning a smaller model or prompting a larger model without fine-tuning it. The Chain-Of-Thought method makes it more difficult to recognize *Entailment* relations and leads to lower F1. As seen above, this mechanically increases Faithfulness. Contrastive Chain-Of-Thought further reduces the number of predicted *Entailment* relations, with an associated increase in Faithfulness. All systems achieve similar Consistency.

The tf-idf baseline was provided by the task organizers. Some of our proposed systems scored below the baseline in some metrics. For instance, Clinical-Longformer obtained a much lower Faithfulness and Consistency, ClinicalBERT, CCOT, and 1SCOT prompts obtained lower F1 scores.

According to the leaderboard, the top scores were 0.80 for the F1 score, 0.95 for Faithfulness, and 0.81 for Consistency, achieved by 3 different teams. We do not have information regarding the approaches these teams chose at the time of writing. Using last year's results on the F1 score, the approach of Kanakarajan and Sankarasubbu (2023), using Flan-T5-xxl, achieved an F1 score of 0.83. Their approach differs from ours by not only prompting Flan-T5 but by finetuning it beforehand using single- and multiple-instruction templates. This approach leads to a boost in performance compared to our simpler approach. Takehana et al. (2023) also performed ensembling and voting of MLMs and achieved an F1 score of 0.66. They performed what we called 'hard voting,' using 10 models for their ensemble and performing data augmentation on the original task dataset. Their result is comparable to our approach using an ensemble of 3 ClinicalBERT or 3 Clinical-Longformer.

5.2 Error analysis

In this section, we analyze our models in more depth by breaking down their results according to gold labels, whether a comparison of CTRs is involved, the types of inference to perform, CTR sections, and examine the F1 score per intervention type. For simplicity, we focus our analysis only on the two best-performing systems of each approach.

Accuracy per gold label From the accuracy displayed in Tab. 3, we observe that our LLM methods, especially CCOT, handle the Contradiction examples better. This label is the most frequent in the test set (67% of instances labeled as Contradiction and 33% as Entailment). MLMs, in contrast, have similar accuracy across both labels.

Comparison versus Single The *Comparison* of 2 CTRs implies longer input sequences and possibly an increased complexity since the model needs to confront the elements of two separate documents. Surprisingly, as reported in Tab. 4, we observe that all models perform similarly for *Comparison* and

System	Entailment	Contradict.
3 NLI-RoBERTa	**55**	56
(d) + Flan-T5-large	**55**	48
2S	44	64
1SCCOT	20	**82**

Table 3: Accuracy (in %) per label: *Entailment* and *Contradiction* (Contradict.). Systems: ensemble of 3 NLI-RoBERTa; ensemble of all MLM baselines (d) + Flan-T5-large (2S); Flan-T5-large in 2-shot (2S) and 1-shot contrastive chain-of-thought (1SCCOT) settings.

Single. We can hypothesize that the models are able to find more clues with 2 documents instead of 1 and predict more accurate labels.

System	Single	Comparison
3 NLI-RoBERTa	56	56
(d) + Flan-T5-large	49	51
2S	59	56
1SCCOT	**61**	**61**

Table 4: Accuracy (in %) per CTR type: *Single* and *Comparison*. Systems: see Tab. 3.

CTR sections From the accuracy displayed in Tab. 5, we observe no performance distinction between the models for different sections.

System	AE	Int.	Elig.	Res.
3 NLI-RoBERTa	60	59	52	52
(d) + Flan-T5-large	43	46	55	57
2S	55	58	**61**	54
1SCCOT	**62**	**63**	58	**60**

Table 5: Accuracy (in %) per CTR section: *Adverse events* (AE), *Intervention* (Int.), *Eligibility* (Elig.), and *Results* (Res.). Systems: see Tab. 3.

Types of 'intervention' Tab. 6 results were obtained directly from the task organizers' evaluation script. Once again NLI-RoBERTa is stable across *Paraphrase* and *Definition* interventions and achieves the best performance. NLI-RoBERTa seems to be less sensitive to semantic change when it comes to paraphrasing. Its score for *Definition* shows that it can capture the relevant information better when more details are provided. Contrastive Chain-Of-Thought does not increase the model's resistance to semantic change (as shown by the results on *Paraphrase*), its ability to perform numerical inference (see results on Numerical paraphrase) or to focus on relevant information (see results on *Definition*). For the latter, the model might struggle to focus on relevant information because of the long length of the input prompts (see Tab. 11).

System	Def.	NP	Para.
3 NLI-RoBERTa	**0.57**	**0.51**	**0.56**
(d) + Flan-T5-large	0.39	0.46	0.54
2S	0.39	0.46	0.54
1SCCOT	0.31	0.26	0.25

Table 6: F1 score per intervention type: *Definition* (Def.), *Numerical Paraphrase* (NP), or *Paraphrase* (Para.) interventions. Systems: see Tab. 3.

6 Conclusion and future work

This paper describes the two systems proposed by the SEME team for the SemEval 2024 Task 2 NLI4CT. Our first approach is based on the finetuning and ensembling of Masked Language Models, using only the challenge's data. Our second approach consists of a pipeline to prompt Large Language Models, using prompt engineering techniques, such as Chain-Of-Thought and Contrastive Chain-of-Thought, in Zero-shot, 1-shot, and 2-shot manners. Our two best-reported results are 0.57 F1 score, 0.64 Faithfulness, and 0.56 Consistency, with prompting Flan-T5-large in a 2-shot manner, ranking 27th out of 32 submissions for F1, 18th for Faithfulness and 25th for Consistency. We obtain the same scores for the MLM system using an ensemble composed of a finetuned NLI-RoBERTa + Clinical-Longformer + ClinicalBERT + the predictions of Flan-T5-large, that is 0.57 for F1 score, 0.64 for Faithfulness, and 0.56 for Consistency.

Some future work could include the continuation of the Masked Language Models pretraining on unlabeled clinical trials, before performing a similar finetuning as presented in the paper. We could also apply this approach to medical Large Language Models like MEDITRON (Chen et al., 2023), by performing instruction-tuning using clinically oriented instructions and then prompting the resulting model on the task data. Another possible approach, similar to (Conceição et al., 2023), would be to incorporate domain ontologies (like UMLS) into the finetuning of Masked Language

Models to provide definitions and supplementary knowledge.

Ethical statement

The NLI4CT task uses clinical data extracted and processed from https://clinicaltrials.gov/. This resource is freely available, provided by the National Library of Medicine, and is an official U.S. Department of Health and Human Services website.

Carbon emissions

Another arguable ethical aspect of our approach is the carbon emissions generated by our models' training and inference. Our experiments used 4 Tesla V100 GPUs paired with 2 Intel Xeon Gold 6148 20 cores and 384 GB of RAM. Depending on the approach chosen, the running time can be up to 10 times longer. For instance, we observe an execution time of 3 hours for the training and inference of an ensemble of 3 ClinicalBERT models. For the inference of Flan-T5-large on a 2-shot Contrastive Chain-Of-Thought, we achieve up to 30 hours of running time to get the predictions for all instances of the test set. Globally, we can say that the MLM approach is computationally more efficient, with running times varying from 3 to 6.5 hours (for the ensemble of ClinicalBERT, NLI-RoBERTa, and Clinical-Longformer). For the LLM approach, we observe running times ranging from 10.5 hours (in Zero-shot) to 38 hours (in 1-shot Chain-Of-Thought).

We used Green Algorithms[8] (Lannelongue et al., 2021) to estimate carbon emissions, taking into consideration our aforementioned computational configuration. The MLM approach produces up to 831g of CO_2 with the 3 models ensembling approach. For the LLM approach, the emissions vary from 1.34 kg of CO_2 for zero, 1, and 2-shot experiments to 4.86kg for Contrastive Chain-Of-Thought experiments.

Considering the little gain in performance of LLMs compared to MLMs using our approach and the CO_2 overconsumption of the LLMs, it would be more reasonable to use the MLM approach in our case. The MLM approach also provides faster predictions, which can be much more convenient.

[8] http://calculator.green-algorithms.org/

Acknowledgements

This work benefited from the GPUs provided by Lab-IA, an institution member of Université Paris-Saclay. This work was also supported through the CNRS grant 80|PRIME.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

Yew Ken Chia, Guizhen Chen, Anh Tuan Luu, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *ArXiv*, abs/2311.09277.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Sofia I. R. Conceição, Diana F. Sousa, Pedro Silvestre, and Francisco M Couto. 2023. lasigeBioTM at SemEval-2023 task 7: Improving natural language inference baseline systems with domain ontologies. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 10–15, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Kamal Raj Kanakarajan and Malaikannan Sankarasubbu. 2023. Saama AI research at SemEval-2023 task 7: Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003, Toronto, Canada. Association for Computational Linguistics.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12).

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu su, and Wenpeng Yin. 2024. MUFFIN: Curating multi-faceted instructions for improving instruction following. In *The Twelfth International Conference on Learning Representations*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bhavish Pahwa and Bhavika Pahwa. 2023. BpHigh at SemEval-2023 task 7: Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Conner Takehana, Dylan Lim, Emirhan Kurtulus, Ramya Iyer, Ellie Tanimura, Pankhuri Aggarwal, Molly Cantillon, Alfred Yu, Sarosh Khan, and Nathan Chi. 2023. Stanford MLab at SemEval 2023 task 7: Neural methods for clinical trial report NLI. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1769–1775, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Hyperparameters

Tab. 7 shows the final hyperparameters used for finetuning the Masked Language Model systems.

Hyperparameter	Value
Nb. epochs	4
Batch size	64
Learning rate	$5e-5$
Optimizer	AdamW

Table 7: Hyperparameters to finetune the MLM systems.

B Example of Natural Language Inference mechanism

Fig. 2 shows an example of the kinds of inference performed by the NLI system in order to predict the correct label.

C Prompts

C.1 Simple prompt

Fig. 3 displays an example Zero-shot prompt. For n-shot prompts, we insert n demonstrations before this prompt. Each demonstration is built from training data; in a demonstration, the Label part is replaced with 'Answer: Yes' or 'Answer: No' depending on whether the example's label is *Entailment* or *Contradiction*.

C.2 Chain-Of-Thought

Fig. 4 displays an example Chain-Of-Thought demonstration. Our initial demonstrations are modified to include the idea of Chain-Of-Thought as mentioned in Wei et al. (2022).

C.3 Contrastive Chain-Of-Thought

Fig. 5 displays an example of our Contrastive Chain-Of-Thought prompt. Our initial demonstrations are modified to include the idea of a Contrastive Chain-Of-Thought as mentioned in Chia et al. (2023).

D NLI4CT dataset statistics

Tab. 8 shows statistics regarding the original task's data, such as the number of CTRs, of statements, the average length of a statement or evidence, and the max length of an evidence or statement.

Metric	Value
Nb. CTRs (documents)	999
Nb. statements	2,400
Avg. length statement	19.5
Max. length statement	65
Avg. length evidence	10.7
Max. length evidence	197

Table 8: Statistics about the NLI4CT train and dev sets.

Subset	*Entailment*	*Contradiction*
Train	850	850
Validation	100	100
Gold test set (whole)	1841	3659
Gold test set (control set)	250	250
Gold test set (contrast set)	1591	3409

Table 9: Statistics about the number of *Contradiction* and *Entailment* instances in NLI4CT dataset.

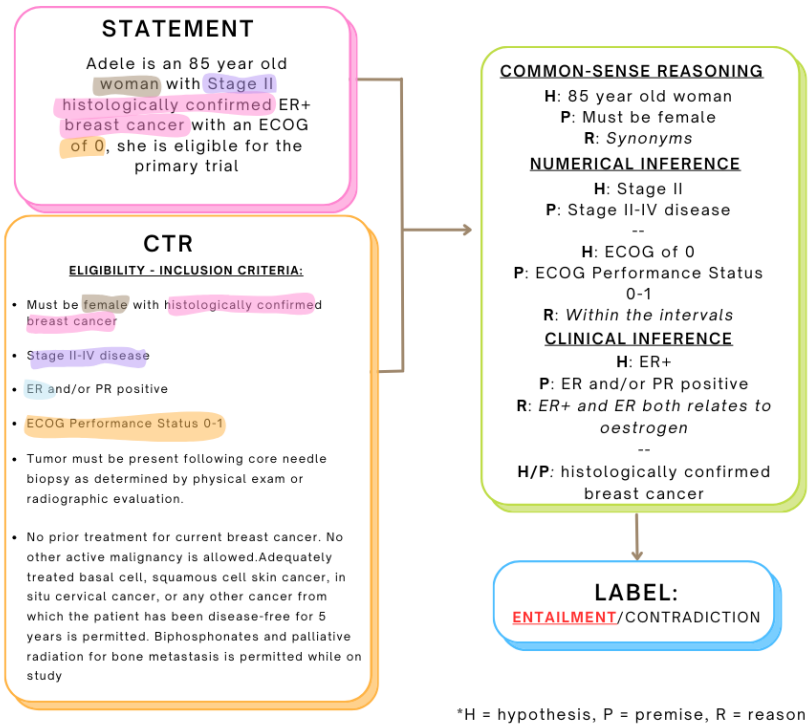

Figure 2: Example of an inference mechanism using a statement and the *Eligibility* section of a CTR.

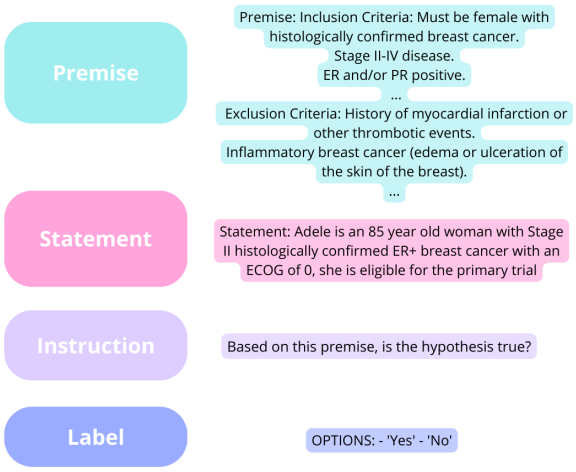

Figure 3: Example Zero-shot prompt.

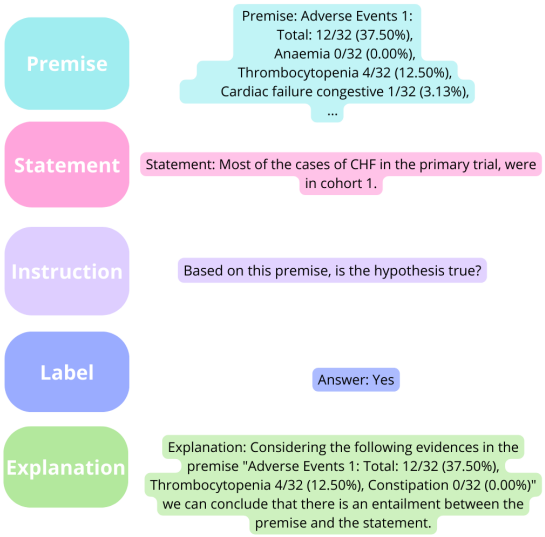

Figure 4: Example Chain-Of-Thought demonstration.

E Metrics on input sequences

E.1 MLM system input sequences

Tab. 10 displays the average, maximum, and minimum length of input sequences for the finetuning of MLMs.

E.2 LLM system input sequences

Tab. 11 displays the average, maximum, and minimum length of prompts used in Flan-T5.

F Prompt selection

Tab. 12 displays the templates tried in order to find the one that would perform the best. The last two prompts were tested using Llama-2 and Mistral. The last prompt uses the concept of 'persona prompting' (Zhang et al., 2018) where we assign the LLM a role.

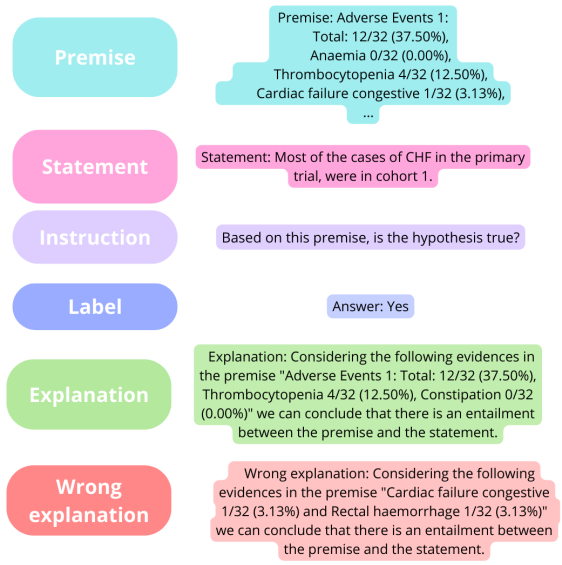

Figure 5: Example Contrastive Chain-Of-Thought demonstration.

Metric	Value
Mean nb. tokens	480
Min. nb. tokens	41
Max. nb. tokens	2799

Table 10: Average, minimum, and maximum number of tokens of an input sequence for the MLM approach.

Prompt	Mean	Min.	Max.
ZS	573	92	1367
1S	1650	835	3009
2S	3036	1397	6669
1S COT	2474	1300	6611
2S COT	3933	6354	2484
1S CCOT	2622	4285	1613
2S CCOT	4826	3153	8321

Table 11: Average, minimum, and maximum numbers of tokens of each kind of prompt for the LLM approach.

Id	Template
1	[Premise] [Statement] Does the premise entail the hypothesis? [Options]
2	[Premise] [Statement] Is the hypothesis entailed by the premise? [Options]
3	[Premise] [Statement] If this premise is true, what does that tell us about whether it entails the hypothesis? [Options]
4	From the following statement and premise, would you say there is a contradiction or an entailment between the statement and the premise? Just answer by saying 'contradiction' or 'entailment'. [Statement] [Premise]
5	Imagine you are a medical practitioner and you are reviewing clinical trials. You are given a statement and a premise. You should determine if there is an entailment or a contradiction between the premise and the statement. There is necessarily an entailment or a contradiction, no neutral case. From the following statement and premise, would you say there is a contradiction or an entailment between the statement and the premise? Just answer by saying 'contradiction' or 'entailment'. [Statement] [Premise]

Table 12: Other prompts tested on the LLM baselines.